Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 5472–5476, June 1992
Biochemistry

A human gene (AHNAK) encoding an unusually large protein with a 1.2-μm polyionic rod structure

(neuroblastoma/protein structure/gene expression)

EMMA SHTIVELMAN*, FRED E. COHEN†, AND J. MICHAEL BISHOP*

*George Williams Hooper Research Foundation, Department of Microbiology and Immunology, and †Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143

Contributed by J. Michael Bishop, January 24, 1992

ABSTRACT **We report here the identification and partial characterization of a human gene (designated AHNAK) that encodes an unusually large protein (≈700 kDa). AHNAK is expressed by means of a 17.5-kilobase mRNA in diverse cellular lineages but is typically repressed in cell lines derived from human neuroblastomas and in several other types of tumors. Unique-sequence domains at the two ends of the protein flank a large internal domain (≈4300 amino acids) composed of highly conserved repeated elements, most of which are 128 amino acids in length. The repeated elements in turn display a redundant motif, marked by the recurrence of proline at every seventh residue. Within these sequences, hydrophobic and hydrophilic residues alternate in a manner that is incompatible with a helical coiled-coil structure. Instead, we propose a structure resembling a β-strand but with a periodicity of 2.33. The structure would engender a polyionic rod ≈1.2 μm long. Preliminary evidence indicates that the protein resides predominantly within the nucleus, but no function has yet been discerned.**

Neuroblastoma represents the most primitive neoplasm originating from migratory neural crest cells and apparently arises as a result of arrested differentiation (1). To identify genes whose transcription might be repressed during the genesis of neuroblastomas, we used (2) subtractive cDNA cloning to detect genes that are expressed in human melanomas and pheochromocytomas but not in neuroblastomas. The first of these genes identified encodes the cell surface protein known as CD44, an integral membrane glycoprotein that is the principal receptor for hyaluronate on the cell surface (3).

A second gene (originally designated PM227) caught our attention because its expression appears to be coordinated with that of CD44 (2). The fact that CD44 and PM227 have similar tissue specificity of expression raised the possibility that both are subject to similar transcriptional regulation and that their downregulation in neuroblastoma is a consequence of the action of the same transcriptional factor(s).

Here we report that PM227 encodes a protein whose exceptionally large size of 700 kDa has caused us to rename the gene AHNAK (meaning giant in Hebrew).‡ The amino acid sequence of the AHNAK protein suggests secondary structure with a periodicity of 2.33 residues per turn. Individual chains could associate to form a seven- or eight-stranded barrel. The resulting structure would be a polyionic rod with length as great as 1.2 μm. The function of these rods remains a mystery.

MATERIALS AND METHODS

Construction of a subtracted cDNA library enriched for sequences expressed in a human pheochromocytoma but not in the neuroblastoma cell line NMB has been described (2). A cDNA library from melanoma cell line HT144 was constructed in phage vector λZAPII according to published protocols (4).

To prepare a genomic minilibrary, high molecular weight DNA from human lymphoblastoid cell line RPMI7666 was digested to completion with *Eco*RI, the restriction fragments were separated on 0.4% agarose gel, and fragments of sizes between 18 and 24–26 kilobase pairs (kbp) were gel-purified and cloned into phage vector λDASH (Stratagene).

Southern and Northern blot hybridizations were performed after transfer of electrophoretically separated nucleic acids to nylon Hybond membranes (Amersham) as described (5).

Templates for DNA sequencing were produced by generating sets of nested deletions of cDNA clones. Nucleotide sequences were determined using the Sequenase kit (United States Biochemical) or automated sequencing performed by the Biomedical Research Facility at the University of California, San Francisco.

RESULTS

Cloning of AHNAK cDNA and Genomic Sequences. We began this work to identify genes whose expression might be specifically diminished or absent in cells of human neuroblastomas. To this end, we constructed and screened a subtracted cDNA library enriched for sequences expressed in a human pheochromocytoma but not in neuroblastomas (2). Briefly, polyadenylylated RNA isolated from a surgical specimen of a benign pheochromocytoma tumor was transcribed into single-stranded cDNA. This cDNA was hybridized with an excess of RNA from the human neuroblastoma cell line NMB. The unhybridized single-stranded cDNA was selected on a hydroxyapatite column as described (6) and back-hybridized to pheochromocytoma RNA to obtain cDNA·RNA hybrids that were used to construct a subtracted cDNA library of 3×10^6 clones in phage vector λgt10. The resulting library was screened with radiolabeled cDNAs prepared with RNAs from pheochromocytoma, neuroblastoma, and melanoma cell lines. Only the clones expressed in both pheochromocytoma and melanoma but not in neuroblastoma were selected. Among the products of this screen was a cDNA clone originally designated as PM227 that we have renamed AHNAK.

The diminished expression of AHNAK RNA in neuroblastoma cell lines is illustrated in Fig. 1. Northern blot analysis of AHNAK RNA demonstrated a transcript estimated to be about 18 kilobases (kb) long in several human cell lines derived from diverse tumors (Fig. 1*A*). The abundance of this transcript was greatly diminished in most of the neuroblastoma lines examined. Because of the very large size of AHNAK RNA, the RNase protection method was used for more extensive analysis of its expression in neuroblastoma

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡The sequences reported in this paper have been deposited in the GenBank data base (accession nos. M80902 and M80899).

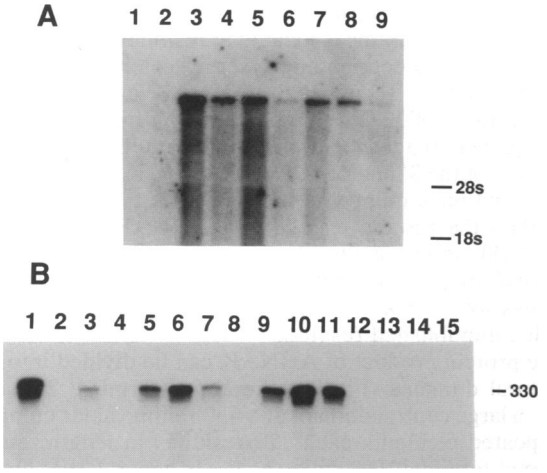

FIG. 1. Expression of AHNAK in human cells. (*A*) Northern blot analysis of AHNAK expression. Oligo(dT)-selected RNA (5 μg) was electrophoresed through a 0.8% agarose gel containing formaldehyde, transferred to a Hybond filter (Amersham), and hybridized to radiolabeled insert of cDNA clone PM227. RNAs from the following human cell lines were analyzed. Lanes: 1, Burkitt lymphoma BL21; 2, colon carcinoma COLO-320HSR; 3, melanoma HT144; 4, promyelocytic leukemia HL-60; 5, osteosarcoma MG63; 6, neuroblastoma NLF; 7, neuroepithelioma SK-N-MC; 8, myosarcoma A204; 9, neuroblastoma NMB. (*B*) RNase protection analysis of AHNAK RNA expression. A RNA probe was made with the 330-nucleotide fragment from the 5′ end of clone z37. Oligo(dT)-selected RNA (2 μg) from human cell lines was annealed with the radiolabeled RNA probe at 47°C and, after digestion with RNases A and T1, the protected products were resolved on a denaturing 5% polyacrylamide gel. RNAs were isolated from the following human cell lines. Lanes: 1, melanoma C32r; 2–8, neuroblastomas MCN-1, NGP, IMR-32, LAN1, LAN2, LAN5, and SK-N-SH, respectively; 9, neuroepithelioma SK-N-MC; 10, promyelocytic leukemia HL-60; 11, surgical specimen of pheochromocytoma; 12 and 13, small cell lung carcinoma cell lines N417 and H82, respectively; 14 and 15, Burkitt lymphoma cell lines Raji and Daudi, respectively.

cell lines (Fig. 1*B*). Among human cell lines analyzed, we have found low or undetectable levels of AHNAK expression in lines derived from neuroblastomas, small cell lung carcinomas, and Burkitt lymphomas. But repression of AHNAK is not a universal feature of neoplasia, since we found expression in all other human cell lines examined.

To clone the full-length cDNA of AHNAK, we used the cDNA clone PM227 as a probe for screening a cDNA library made with polyadenylylated RNA from the human melanoma cell line HT144. Five cDNA clones (z2, z6, z7, z8, and z18) were obtained, mapped, and shown to span 4 kb of AHNAK RNA (Fig. 2). RNase protection analyses performed with RNA probes synthesized from both strands of a cDNA fragment were used to establish the transcriptional orientation of cDNAs (data not shown). Clone z18 was found to extend farthest in the 5′ direction (Fig. 2). We performed a "cDNA walking" experiment to isolate AHNAK cDNA clones representing the middle portion and the 5′ end of AHNAK RNA. A 5′-terminal fragment of clone z18 was used as a probe for repeated screening of the HT144 cDNA library. This round of screening produced 10 cDNA clones, of which the four longest (z21, z78, z80, and z83) are shown in Fig. 2.

Analysis of these cDNA clones revealed that all are composed of cross-hybridizing sequences. The only unique sequences were contained within 1.5 kb at the 5′ terminus of clone z21 and 3.3 kb at the 3′ end of cDNA clone z7. Repeated screening of the cDNA library with the unique probes did not produce additional clones extending farther in the 3′ direction; one clone, z37, was identified with the 5′-end-specific probe and was found to extend only slightly farther in the 5′ direction than cDNA z21.

Restriction mapping and partial sequencing of cDNA clones allowed us to map clone z80 in relation to the isolated cDNAs (Fig. 2). However, we could not find any regions of overlap among cDNA clones z21 and z83 and other clones. We used the cDNA clones as probes in Southern and Northern blot analysis to prove that they are all derived from a single genomic locus and transcript rather than from several different though homologous genes. We have found that all cDNAs examined detect the same size RNA species in Northern blot analysis (data not shown). Southern blot analysis showed that all cDNAs hybridize to an overlapping set of restriction fragments in genomic DNA and, more importantly, single restriction fragments were detected in *Eco*RI and *Kpn* I digests (data not shown). These data indicated that all cDNA clones of the z series are derived from a single genomic locus.

To isolate and characterize the genomic locus of AHNAK, two phage clones, λz3 and λz8, were isolated from a normal human placental genomic library with cDNA probes (Fig. 2). Both clones contained sequences corresponding to the 5′ end of the cDNAs z21 and z37 only. To obtain genomic sequences coding for the remainder of the transcript represented in AHNAK cDNAs, we took advantage of the fact that a single *Eco*RI fragment of 24 kbp was detected on Southern blots with all cDNA clones except z21 and z37, which have an internal *Eco*RI site close to their 5′ ends and detect an additional 5.2-kbp *Eco*RI fragment. Genomic *Eco*RI fragments of about 24 kbp were cloned into vector λDASH. Screening of the resulting minilibrary produced one positive clone, designated λRR00. The insert of this clone was only 16 kbp long instead of the expected 24 kbp. We have found that the only rearrangement suffered by clone λRR00 is a 3′ terminal deletion of 8 kbp. This was proved through detailed restriction analysis of λRR00, comparison of its restriction map with that derived from genomic blot analysis, comparison with the cDNA maps, and partial sequencing of the genomic subclones and cDNAs.

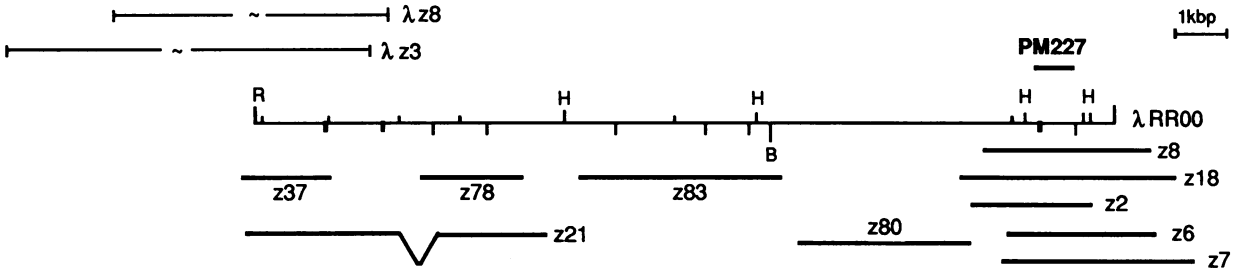

FIG. 2. Schematic representation of the genomic locus and cDNAs of AHNAK. Genomic clones, thin lines; cDNA clones, thick lines. Clones λz3 and λz8 are not shown to scale. Only some of the restriction sites used in the mapping of the genomic and cDNA clones are marked on the λRR00 map. R, *Eco*RI; H, *Hin*dIII; B, *Bam*HI; thin vertical bars above the line representing λRR00, *Eco*RV; thin vertical bars beneath the line, *Nsi* I; thick vertical bars, *Nde* I.

Fig. 2 shows the alignment of AHNAK cDNAs and genomic clones. All cDNA clones are colinear with the genomic sequences with the exception of clone z21, which has a deletion of about 0.8 kbp when compared with the corresponding genomic region. However, cDNA z78 contains genomic sequences missing in z21. Subsequent sequence analyses of cDNAs and genomic subclones indicated that clone z21 most likely suffered an internal deletion as a cloning artifact. This conclusion was supported by two findings: (*i*) the genomic region of 768 nucleotides absent from cDNA z21 is not flanked by sequences even remotely resembling splicing signals and (*ii*) the open reading frame of AHNAK continues uninterrupted through this fragment with the repetitive nature of the predicted amino acid sequence preserved (see below). Thus, with the exception of two small gaps between cDNA clones z21 and z83 and between z83 and z80 (Fig. 2), all of the genomic sequences are represented in cDNA clones.

The gaps between cDNAs and genomic sequences most likely reflect deficiencies in the cDNA library rather than regions of splicing. This deduction was based on the presence of an uninterrupted open reading frame and absence of splicing signals in the two genomic regions not represented in cDNA. Thus, we suggest that the cloned portion of the AHNAK genomic locus has no introns, which is remarkable considering the size of the AHNAK transcript. A 3′ domain of 1.4 kbp represented in cDNAs is not present in the genomic clone, but genomic Southern blot analysis indicates that it is contained in a contiguous genomic fragment.

We examined next the phylogenetic conservation of AHNAK. Fig. 3 shows that cDNA clone z80 detects bands in the DNAs of bovine, rodent, and quail cells. We were also able to detect hybridizing fragments in *Drosophila* and *Dictyostelium discoideum* DNA under conditions of reduced stringency (data not shown). In addition, AHNAK cDNAs detected an ≈18-kb transcript in RNA from rodent fibroblasts (data not shown).

Analysis of the AHNAK Nucleotide Sequence and Predicted Amino Acid Sequence. By compiling nucleotide sequences from both cDNA and genomic clones, we assembled contiguous sequences of 5.5 and 3.8 kbp from the 5′ and 3′ ends of AHNAK, respectively. Further analysis revealed that cDNA clone z37 extends to the authentic 5′ end of AHNAK RNA (data not shown). An open reading frame begins at nucleotides 404–406, with an AUG preceded by a consensus sequence favorable for initiation of translation (CACCATG)

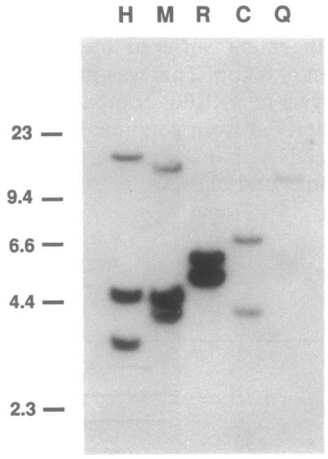

FIG. 3. Phylogenetic conservation of AHNAK sequences. Genomic DNAs from human (H), bovine (C), rat (R), mouse (M), and quail (Q) cells were digested with *Hin*dIII and subjected to Southern blot analysis with the cDNA insert of clone z80 as a probe (Fig. 2). Molecular sizes in kb are indicated.

(7). A second AUG with a similar consensus sequence and in the same reading frame is located at residues 587–589. Either or both of these codons might serve to initiate translation. The available 3′ sequence of AHNAK contains an open reading frame of 1277 codons, extending from the 5′ end of the sequence to a termination codon located 219 nucleotides upstream of the 3′ end of cDNA clone z7. If the open reading frame continues uninterrupted throughout the AHNAK locus, then the protein encoded by AHNAK contains more than 5600 amino acids. In accord with this deduction, AHNAK-specific antibodies detect a protein with a molecular mass of ≈700 kDa in cells expressing the gene (E.S. and J.M.B., unpublished results).

The protein product of AHNAK can be divided into three structural domains (Fig. 4): the amino-terminal 251 amino acids, a large central domain of ≈4300 amino acids composed of repeated elements each 128 residues in length, and the carboxyl-terminal 1002 amino acids. In general, the sequence is rich in apolar and charged residues. The middle portion contains an abundance of both negatively and positively charged residues (pI 6.5), whereas both terminal domains are rich in lysines and, thus, are basic (pI 8.8). In addition, both terminal domains are rich in glycine.

The repeated elements that compose the central domain of AHNAK protein are highly conserved: identity between any two units is on average 80%, and virtually all substitutions are conservative, preserving polarity and charge (Fig. 4). The complete unit of repetition appears to be 128 amino acids long, but some of the units are reduced by deletions of 7, 54, or 61 residues, located at the same position within the repeat (Fig. 4). In contrast, the most amino-terminal of the repeats has an insertion of 69 amino acids.

Searches of the protein sequence data base (Protein Identification Resource, January 1992) revealed statistically significant homology only with members of collagen and elastin families. This results from the fact that all these sequences are rich in glycine and proline and are repetitive in nature. Inspection of the AHNAK amino acid sequence revealed the presence in the carboxyl-terminal domain of three copies of the sequence Lys-Ser-Pro-Lys, which is identical to the cdc2 kinase phosphorylation sites in H1 histone (8, 9). The carboxyl terminus displays several stretches with a high density of positively charged residues (see Fig. 6) reminiscent of nuclear localization signals (10).

A Structural Model of AHNAK Protein. Inspection of the 128-residue repeat revealed an underlying heptad repeat with the motif $\varphi\pm\varphi P\pm\varphi\pm$, where φ is a hydrophobic residue, $\pm$ is a hydrophilic residue that is often charged, and P is proline (Fig. 5). Variations on this theme incorporate insertions or deletions of 2 or rarely 4 residues. This preserves the alternating character of the sequence. A similar pattern of heptad repeats is found also in the middle portion of the unique part of the sequenced carboxyl-terminal domain (amino acids 474–635, Fig. 4 *Lower*), though this region of the protein does not display the characteristic 128-amino acid repeats. Heptad repeats are common in fibrous proteins such as tropomyosin and keratin (11) and in nuclear proteins containing leucine zippers (12). α-Helical structure with 3.5 residues per turn has been demonstrated in these cases (13, 14). Prolines are rarely observed in these sequences. The carboxyl terminus of RNA polymerase II contains two prolines at positions 3 and 6 within a heptad repeat (15) and is presumed to form a pair of overlapping β-turns. The AHNAK heptad is incompatible with either of these structures. Instead, we propose a structure that accommodates the AHNAK heptad repeat.

A Fourier analysis (16) of the hydrophobicity of the entire sequence contains a dominant feature with a period of 2.33 residues per cycle. Peptides adopt a repeating structure with 2.33 residues per cycle when the backbone dihedral angles satisfy the relation $/\phi+\psi/ = 25$. When $\phi+\psi > 0$, the

```
MPGIKVGGSGVNVNAKGLDLGGRGGVQVPAVDISSSLGGRPVEVQGPSLESGDHAKIKFPTMKVPKFGVSTGREGQTPKAGLRVSAPEVSVGHKGGKPGL
TIQAPQLEVSVPSANIEGLEGKLKGPQITGPSLEGDLGLKGAKPQGHIGVDASAPQIGGSITGPSVEVQAPDIDVQGPGSKLNVPKMKVPKFSVSGAKGE
ETGIDVTLPTGEVTVPGVSGDVSLPEIATGGLEGKMKGTKVKTPEMIIQKP                                                  251
KISMQDVDLSLGSPKLKGD IKVSAPGVQGDVKGPQVALKGSRVDIETPNLEGTLTGPRLGSPSGKTGTCRISMSEVDLNVAAPKVKGG
              VDVTLPRVEGKVKVPEVDVRGPKVDVSAPDVEAHGPEWNLKMPKMKMPTFSTPGAKGEGPDVHMTLPKGDISISGPKVNVEAPDVNLEGLGGKLKGPDVKLPDMSVKTP  448
KISMPDVDLHVKGTKVKGEYDVTVPKLEGELKGP       KVDIDAPDVDVHGPDWHLKMPKMKMPKFSVPGFKAEGPEVDVNLPKADVDISGPKIDVTAPDVSIEEPEGKLKGPKFKMPEMNIKVP  569
KISMPDVDLHLKGPNVKGEYDVTMPKVESEIKVPDVELKSAKMDIDVPDVEVQGPDWHLKMPKMKMPKFSMPGFKAEGPEVDVNLPKADVDISGPKVGVEVPDVNIEGPEGKLKGPKFKMPEMNIKAP  697
KISMPDVDLHMKGPKVKGEYDMTVPKLEGDLKGP       KVDVSAPDVEMQGPDWNLKMPKIKMPKFSMPSLKGEGPEFDVNLSKANVDISAPKVDTNAPDLSLEGPEGKLKGPKFKMPEMHFRAP  818
KMSLPDVDLDLKGPKMKGNVDISAPKIEGEMQVPDVDIRGPKVDIKAPDVEGQGLDWSLKIPKMKMPKFSMPSLKGEGPEVDVNLPKADVDVSGPKVDIEAPDVSLEGPEGKLKGPKFKMPEMHFKTP  946
KISMPDVDLHLKGPKVKGDVDVSVPKVEGEMKVPDVEIKGPKMDIDAPDVEVQGPDWHLKMPKMKMPKFSMPGFKGEGREVDVNLPKADIDVSGPKVDVEVPDVSLEGPEGKLKGPKFKMPEMHFKAP 1074
KISMPDVDLNLKGPKLKGDVDVSLPEVEGEMKVPDVDIKGPKVDISAPDVDVHGPDWHLKMPKVKMPKFSMPGFKGEGPEVDVKLPKADVDVSGPKMDAEVPDVNIEGPDAKLKGPKFKMPEMSIKPQ 1202
KISIPDVGLHLKGPKMKGDYDVTVPKVEGEIKAPDVDIKGPKVDINAPDVEVHGPDWHLKMPKVKMPKFSMPGFKGEGPEVDMNLPKADLGVSGPKVDIDVPDVNLEAPEGKLKGPKFKMPSMNIQTH 1330
KISMPDVGLNLKAPKLKTDVDVSLPKVEGDLKGP                                                  EIDVKAPKMDVNVGDIDIEGPEGKLKGPKFKMPEMHFKAP 1404
KISMPDVDLHLKGPKVKGDMDVSVPKVEGEMKVPDVDIKGPKVDIDAPDVEVHDPDWHLKMPKMKMPKFSMPGFKAEGPEVDVNLRKADIDVSGPSVDTDAPDLDIEGPEGKLKGSKFKMPKLNIKAP 1532
KVSMPDVDLNLKGPKLKGEIDASVPELEGDLRGP                                                  QVDVKGPLVEAEVPDVDLECPDAKLKGPKFKMPEMHFKAP 1606
KISMPDVDLHLKGPKVKGDADVSVPKLEGDLTGP                                                  SVGVEVPDVELECPDAKLKGPKFKMPDMHFKAP 1673
KISMPDVDLH                                                                                                                    1683
```

```
                    TLPKVEGDLKGPEADIKGPKVDINTPDVDVHGPDWHLKMPKVKMPKFSMPGFKGEGPDVDVNLPKADIDVSGPKVDVDVPDVNIEGPDAKLKGPKFKMPEMSIKAP  106
KISMPDIDLNLKGPKVKGDVDVTLPKVEGDLKGPEADIKGPKVDINTPDVDVHGPDWHLKMPKVKMPKFSMPGFKGEGPDVDVSLPKADIDVSGPKVDVDIPDVNIEGPDPKLKVPKVKMPEINIKAP  234
KISIPDVDLDLKGPKVKGDFDVSVPKVEGTLKGPEVDLKGP                                                                                        265
RLDFEGPDAKLSGPSLKMPSLEISAPKVTAPDVDLHLKAPKIGFSGPKLEGGEVDLKGPKVEAPSLDVHMDSPDINIEGPDVKIPKFKKPKFGFGPKSPK                          365
ADIKSPSLDVTVPEAELNLETPEISVGGKGKKSKFKMPKIHMSGPKIKAKKQGFDLNVPGGEIDASLKAPDVDVNIAGPDAALKVDVKSPKTKKTMFGKM                          465
YFPDVEFDIKSPKFKAEAPLPSPKLEGELQAPDLELSLPAIHVEGLDIKAKAPKVKMPDVDISVPKIEGDLKGPKVQAHLGAPDINIEGLDAKVKTPSFG                          565
ISAPQVSIPDVNVNLKGPKIKGDVPSVGLEGPDVDLQGPEAKIKFPKFSMPKIGIPGVKMEGGGAEVHAQLPSLEGDLRGPDVKLEGPDVSLKGPGVDLP                          665
SVNLSMPKVSGPDLDLNLKGPSLKGDLDASVPSMKVHAPGLNLSGVGGKMQVGGDGVKVPGIDATTKLNVGAPDVTLRGPSLQGDLAVSGDIKCPKVSVG                          765
APDLSLEASEGSIKLPKMKLPQFGISTPGSDLHVHAKGPQVSGELKGPGVDVNLKGSRISAPNVDFNLEGPKVKGSLGATGEIKGPTVGGGLPGIGVQGL                          865
EGNLQMPGIKSSGCDVNLPGVNVKLPTGQISGPEIKGGLKGSEVGFHGAAPDISVKGPAFNMASPESDFGINLKGPKIKGGADVSGGVSAPDISLGEGHL                          965
SVKGSGGEWKGPQVSSALNLDTSKFAGGLHFSGPKVEGGVKGGQIGLQAPGLSVSGPQGHLESGSGKVTFPKMKIPKFTFSGRELVGREMGVDVHFPKAE                         1065
ASIQAGAGDGEWEESEVKLKKSKIKMPKFNFSKPKGKGGVTGSPEASISGSKGDLKSSKASLGSLEGEAEAEASSPKGKFSLFKSKKPRHRSNSFSDERE                         1165
FSGPSTPTGTLEFEGGEVSLEGGKVKGKHGKLKFGTFGGLGSKSKGHYEVTGSDDETGKLQGSGVSLASKKSRLSSSSSNDSGNKVGIQLPEVELSVSTKKE                         1277
```

FIG. 4. Predicted partial amino acid sequence of AHNAK. (*Upper*) Amino-terminal 1683 amino acids. (*Lower*) Carboxyl-terminal 1277 residues. Sequences of the amino and carboxyl termini of the protein are in boldface type; the repeats constituting the middle domain are aligned; the insert in the first repeat is shown in italic type. Underlined residues are the putative nuclear localization signals, and the tetrapeptides KSPK at the carboxyl terminus.

structure is left-handed. The pyrrolidone ring of proline restricts the backbone ϕ angle to values near $-75°$ to $-85°$. If the backbone dihedral angles throughout the heptad are similar, then $\phi = -85°$ and $\psi = +110°$ or $+60°$. An examination of a Ramachandran plot of peptide energetics as a function of backbone dihedral angles suggests that $\phi = -85°$ and $\psi = +110°$ is a more stable structure (17). A molecular model of this structure was constructed and regularized using energy minimization with the AMBER force field. The average ϕ,ψ angles after minimization are $(-80°, +109°)$.

Fig. 6 illustrates a molecular model of the proposed structure. The model cleanly separates the hydrophobic and hydrophilic groups in space and only a small distortion is necessary to accommodate the proline residues. The structure is similar to a β-strand, suggesting that the chains could aggregate and form a barrel structure with intermolecular hydrogen bonds. A seven-stranded barrel with a diameter of 9.8 Å or an eight-stranded barrel with a diameter of 10.1 Å would satisfy the geometric requirements for hydrogen bonding. Hydrophobic groups would reside in the interior exclusively, and charged and other hydrophilic groups would cover the exterior. Seven-stranded structures have been

```
φ±φP±φ±

ISMPDIDLN
LKGPKVKGDVD
VTLPKVEGD
LKGPEAD
IKGPKVD
INTPDVD
VHGPDWH
LKMPK
VKMPK
FSMPGFK
GEGPDVD
VSLPKADID
VSGPKVD
VDIPDVN
IEGPDPK
LKVPK
VKMPEIN
IKAPK
```

FIG. 5. Heptad repeats in AHNAK protein. The sequence of the 128-amino acid repeat from the carboxyl-terminal portion of AHNAK sequence (Fig. 4, residues 108–235) highlights the underlying heptad repeat and the conservation of proline every seven residues.

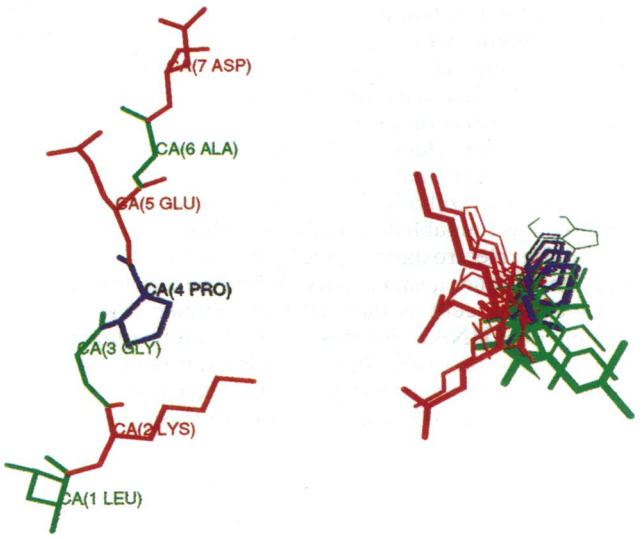

FIG. 6. Two views of the predicted conformation of the repeating structure of the AHNAK heptad. Hydrophilic residues are shown in red, hydrophobic residues are in green, and proline residues are in blue. (*Left*) The extended geometry of the polypeptide backbone (sequence of heptad 4 shown in Fig. 5 was used). (*Right*) An end-on view of repeating units (heptads 4–7 in Fig. 5) demonstrating the segregation of hydrophobic (alanine, valine, leucine, isoleucine, and tryptophan) and hydrophilic (lysine, aspartic acid, and glutamic acid) residues.

described at the domain interface of globular proteins (18) and eight-stranded barrels are common components of globular proteins (19). In either case, the resulting structure would be a polyionic rod as long as 1.2 μm.

DISCUSSION

We have described here the identification of a human gene, AHNAK, cloned initially as one of the genes whose expression is down-modulated in neoplastic neuroblasts, in comparison to differentiated derivatives of the neuroectodermal lineage. As reported (2), another member of this differentially expressed group of genes was CD44. The glycoprotein product of CD44 mediates many functions in a variety of cell types (20). CD44 appears to be expressed on all nonneuronal derivatives of neural crest examined so far (21) with the exception of neuroblastoma cells (2). We have speculated that downregulation of CD44 in neuroblastoma cells might be relevant to the arrested differentiation of neuroblastoma cells or to the highly metastatic properties of this tumor (2). The expression pattern of AHNAK in human cells closely parallels that of CD44, which prompted us to initiate analysis of AHNAK structure and regulation.

AHNAK is a human gene coding for an unusually large protein with a highly periodic amino acid sequence and a predicted secondary structure that would be unprecedented, to our knowledge, if correct. We have found several salient structural features during analysis of AHNAK genomic sequences and the predicted amino acid sequences. (*i*) The genomic locus of AHNAK most likely lacks introns. This is an unexpected finding considering the large size of the AHNAK transcript, second only to that of the muscle-specific gene titin (22). (*ii*) A sequence of $\approx$4400 amino acids in the middle part of the AHNAK protein is composed of multiple repeats of a 128-residue motif. Repetitive structures were described for several other very large protein products, including titin (23), twitchin (24), dystrophin (25), and *Paramecium primaurelia* G surface protein (26). All these large proteins share a general three-domain structure with the repeated motifs comprising the middle portion of the proteins. A distinctive feature of the AHNAK repeats is the high degree of amino acid sequence identity between them. (*iii*) Each 128-amino acid repeat in turn exhibits an internal repetitive structure apparent as heptad repeats with proline at almost every seventh position.

There are few clues to the function of AHNAK. The rod-like structure that we predict for the gene product suggests a structural role. The protein appears to be localized in the nucleus (unpublished analyses), where it could serve as part of the electrostatic "glue" that helps the polyionic components share a small space. When the differentiation of neuroblastoma cells is induced with retinoic acid, the expression of AHNAK increases coordinately (unpublished data). Thus, it is possible that the AHNAK protein is required for the differentiated phenotype of neuronal cells. Since our data suggest that AHNAK may not be expressed in all types of cells, the function of the gene may be specialized rather than universal.

We thank R. Goold of the Biomedical Research Center at the University of California, San Francisco, for assistance in DNA sequencing. This work was supported by grants from the National Institutes of Health (CA 44338) and by funds from the George Williams Hooper Research Foundation.

1. Helman, L. J., Thiele, C. J., Linehan, W. M. & Israel, M. A. (1986) in *Biochemistry and Molecular Genetics of Cancer Metastasis*, eds. Lapis, K., Liotta, L. A. & Rabson, A. S. (Nijhoff, Boston), pp. 73–82.
2. Shtivelman, E. & Bishop, J. M. (1991) *Mol. Cell. Biol.* **11**, 5446–5453.
3. Aruffo, A., Stamencovic, I., Melnick, M., Underhill, C. B. & Seed, B. (1990) *Cell* **61**, 1303–1313.
4. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
5. Church, G. M. & Gilbert, W. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 1991–1995.
6. Hedrick, S. M., Cohen, D. I., Nielsen, E. A. & Davis, M. M. (1984) *Nature (London)* **308**, 149–153.
7. Kozak, M. (1986) *Cell* **44**, 283–292.
8. Langan, T. A., Zeilig, C. E. & Leichtling, B. (1980) in *Protein Phosphorylation and Bio-Regulation* (Karger, Basel), pp. 70–82.
9. Langan, T. A., Gautier, J., Lohka, M., Hollingsworth, R., Moreno, S., Nurse, P., Maller, J. & Sclafani, R. A. (1989) *Mol. Cell. Biol.* **9**, 3860–3868.
10. Dingwall, C. & Laskey, R. A. (1986) *Annu. Rev. Cell Biol.* **2**, 367–390.
11. McLachlan, A. D. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52**, 411–420.
12. Landshulz, W. H., Johnson, P. F. & McKnight, S. L. (1988) *Science* **240**, 1759–1769.
13. McLachlan, A. D. & Karn, J. (1982) *Nature (London)* **299**, 266–269.
14. Rasmussen, R., Benvegnu, D., O'Shea, E. K., Kim, P. S. & Alber, T. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 561–564.
15. Suzuki, M. (1990) *Nature (London)* **344**, 562–565.
16. Cornete, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. & DeLisi, C. (1987) *J. Mol. Biol.* **195**, 659–685.
17. Anderson, A. G. & Hermans, J. (1988) *Proteins* **3**, 262–265.
18. Murzin, A. & Chothia, C. E. (1992) *J. Mol. Biol.* **223**, 531–543.
19. Chothia, C. E. & Finkelstein, A. V. (1990) *Annu. Rev. Biochem.* **59**, 1007–1039.
20. Haynes, B. F., Liao, H.-X. & Patton, K. (1991) *Cancer Cells* **3**, 347–350.
21. Picker, L. J., Nakache, M. & Butcher, E. C. (1989) *J. Cell Biol.* **109**, 927–937.
22. Kurzban, G. P. & Wang, K. (1988) *Biochem. Biophys. Res. Commun.* **155**, 1155–1161.
23. Labeit, S., Barlow, D. P., Gautel, M., Gibson, T., Holt, J., Hsieh, C.-L., Francke, U., Leonard, K., Wardale, J., Whiting, A. & Trinick, J. (1990) *Nature (London)* **345**, 273–276.
24. Benian, G. M., Kiff, J. E., Neckelmann, N., Moerman, D. G. & Waterston, R. H. (1989) *Nature (London)* **342**, 45–50.
25. Hoffman, E. P., Brown, R. H. & Kunkel, L. M. (1987) *Cell* **51**, 919–928.
26. Prat, A., Katinka, M., Caron, F. & Meyer, E. (1986) *J. Mol. Biol.* **189**, 47–60.